LA-UR-04-2239

Approved for Public Release
Distribution is Unlimited

Title: **An Empirical Performance Analysis of Commodity Memories in Commodity Servers**

Authors: **Darren J. Kerbyson, CCS-3**

Mike Lang, CCS-3

Gene Patino, SMART Modular Technologies

Hossein Amidi, SMART Modular Technologies

Published: **in proc. of ACM workshop on Memory System Performance (MSP'04), Washington DC, June 2004.**

CCS-3 REPRINT
Modeling, Algorithms, and Informatics Group

Performance and Architecture Lab
http://www.c3.lanl.gov/par_arch

An Empirical Performance Analysis of Commodity Memories in Commodity Servers.

Darren J. Kerbyson, Mike Lang
Performance and Architecture Lab (PAL)
Los Alamos National Laboratory
NM 87545
+1 (505) 667 4913
{djk,mlang}@lanl.gov

Gene Patino, Hossein Amidi
SMART Modular Technologies Inc.
15635 Alton Parkway
Irvine, CA 92618
+1 (949) 753 0116 ext. 129, 127
{Gene.Patino,Hossein.Amidi}@smartm.com

ABSTRACT
This work details a performance study of six different types of commodity memories in two commodity server nodes. A number of micro-benchmarks are used that measure low-level performance characteristics, as well as two applications representative of the ASC workload. The memories vary both in terms of performance, including latency and bandwidths, and in terms of their physical properties and manufacturer. The two server nodes analyzed were an Itanium-II Madison based system, and a Xeon based system. All memories can be used within both of these processing nodes. This allows the performance of the memories to be directly examined while keeping all other factors within a node the same (processor, motherboard, operating system etc.). The results of this study show that there can be a significant difference in application performance depending on the actual memory used – by as much as 20%. The achieved performance is a result of the integration of the memory into the node as well as how the applications actually utilize it.

Categories and Subject Descriptors
C.4 [**Performance of Systems**]: Measurement techniques and Performance attributes.

General Terms
Performance

Keywords
Memory System Performance, Memory Modules, Performance Measurement, Performance Analysis.

1. INTRODUCTION
System memory is typically a large component in the cost of any computer purchase. However, its specification is usually distilled down to just a size in giga-bytes per-processor or per-node. This investigation shows that a closer look at memory is necessary in order to achieve a higher overall system performance. There has been much in the way of performance analysis of commodity based cluster systems but little that directly analyzes the differences in commodity memories. The differential in cost between the memories is minimal while, as we will show, the performance differential can amount to 10's of percent.

The characteristics of commodity memories include:

Bandwidth – the bus speed of the processing node determines the bandwidth that a memory module can operate at – this is fixed for a particular chipset within a node.

Latency – the latency to the memory from the processing node is a significant performance factor. There are several parts to latency including: CL latency (or CAS Column-Address-Strobe Latency), the Row Precharge Time (tRP), and the Row Address to Column Address Delay (tRCD).

Packaging – memory modules vary in physical dimensions, and in DRAM IC packaging including TSOP (Thin Small Outline Package) and BGA (Ball Grid Array) packages.

Manufacturer – The actual manufacturing process used can result in different performance tolerances.

When considering the performance of a memory, the module is typically referred to by its bus speed (e.g. PC2100 or 266MHz), CL latency, tRP and tRCD. For example a PC2100 CL2.0-2-2 memory works on a 266MHz bus and has a CL latency of 2.0 cycles with a tRP of 2 cycles and a tRCD of 2 cycles. One would expect that the higher performing memories have a higher rated bandwidth and lower rated latency. As we will show, this is not necessarily the case.

Two nodes are used in this work – a Dell PowerEdge 2650 server containing two Intel 2.8-GHz Xeon processors and a Dell PowerEdge 3250 server containing two Itanium-II 1.3-GHz Madison processors. Both are commonly used in the construction of high-performance clusters. The Dell 2650 uses the ServerWorks GC-LE chipset, and the Dell 3250 uses the Intel E8870 chipset.

The memory modules that were made available for testing are listed in Table 1. The memories are ordered in terms of their CL-tRP-tRCD latencies. As can be seen, the memories also differ in terms of their manufacturer, packaging, and physical dimensions. The first four memory modules were supplied by Smart Modular Technologies, module 5 was obtained in the purchase of the Dell 2650, and module 6 was obtained in the purchase of the Dell 3250. The capacity of all the memory modules was 1GB of which four were used at a time in the testing. All memories had a rating of PC2100 or 266MHz.

MSP'04, June 8, 2004, Washington, DC, USA.

Table 1. Characteristics of the memory modules analyzed.

	Module Part#	Package	CAS Latency	Chip Maker	Height
1	SM12872RDDR3222L	TSOP	CL2.0-2-2	Infineon	1.2"
2	SM12872RDDR301BG-I	BGA	CL2.0-3-3	Infineon	1.125"
3	SM12872RDDR301LP-I	TSOP	CL2.0-3-3	Infineon	1.2"
4	SM12872RDDR301BGAS	BGA	CL2.0-3-3	Samsung	1.2”
5	NL9127RD64042-D21J	TSOP	CL2.5-3-3	Nanya	2.0"
6	MT36VDDT12872G-265C2	TSOP	CL2.5-3-3	Micron	1.2”

The performance characteristics of a memory module are defined by the SPD (Serial Presence Detect) – a 128-byte EEPROM [3] which exists on every module. Relevant characteristics are listed in Table 2 for the 6 memory modules listed in Table 1. The SPD allows auto-configuration of the memory between the motherboard and the module. It defines the time delays that the memory module requires in order to correctly function in the system in which it is installed. If different modules are placed within a node, the memory performance will typically be that of the slowest module.

Table 2. Performance characteristics of the memory modules.

Characteristic	1	2	3	4	5	6
Cycle Time (SDRAM) highest CAS latency (ns)	7	7.5	7.5	7.5	7.5	7.5
Min. Row Precharge Time (ns)	15	20	20	20	20	20
Min. Row Active to Row Active (ns)	15	15	15	15	15	15
RAS to CAS delay (ns)	15	20	20	20	20	20

Previous analysis of commodity memory performance has been confined to optimizing the performance of individual desktop machines, or in relation to over-clocking and maximizing the performance for gaming applications. The authors are not aware of other in-depth performance analysis of commodity memories on scientific workloads.

In this analysis we examine the performance of commodity memories on several application codes that are representative of the ASC (Accelerated Strategic Computing) workload. These include Sweep3D – a deterministic particle transport code [1], and SAGE – a hydro Adaptive Mesh Refinement (AMR) code [4]. Both are scientific applications. In addition micro-benchmarks are utilized that measure the low-level characteristics of memory latency and memory bandwidth. In all cases the performance was measured after a clean reboot on each of the two nodes.

Although the results quantify the performance impact of the different memories on these applications and processing nodes, a qualitative result is that the performance can vary by percentage points (or higher) from a change in the memory. The difference in performance is dependent on several factors including: the make-up of the workload that is processed, the chipset within the processing node, and the memory module. Thus by careful examination of the performance impact different memories have on a workload, a higher level of performance may be achievable with a minimal increase in cost per node.

2. OVERVIEW OF THE PROCESSING NODES

2.1 The Dell 2650 dual Xeon Server

The Dell 2650 dual Xeon Server is constructed using the Grand Champion LE (GC-LE) chipset. This can support up to two Pentium 4 Xeon processors. The GC-LE integrates the functions of a dual channel DDR memory controller with interfaces to two PCI-X I/O controllers which can operate at 3.2GB/s as shown in Figure 1. It can support up to 16GB of main memory at speeds up to 4.2GB/s bandwidth. The processor bus is 8 bytes wide operating at a frequency of 100MHz and is quad-pumped.

Each Xeon processor is clocked at 2.8GHz. Each processor has a 12KB L1 instruction cache, an 8KB L1 data cache, and a 512KB 8-way set-associative L2 cache.

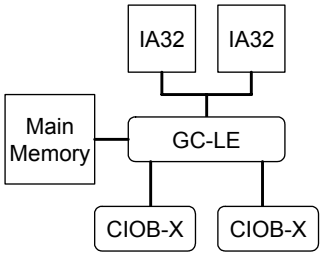

Figure 1. The configuration of a 2-way Xeon Node using the GC-LE chipset.

2.2 The Dell 3250 dual Itanium-II Server

The Dell Itanium-II server is constructed using the Intel E8870 chipset [2]. This chipset enables up to four processors to be placed in a single node, and up to 16 processors in a multi-node configuration using the Scalability Port Switch (SPS) as shown in Figure 2. Each Scalable Node Controller (SNC) supports up to four processors and 2 DDR channels. Up to four memory DIMMS may be places on each channel. A single system bus connects the four processors to the SNC. This is 16 bytes wide and is clocked at 200MHz. It is double pumped resulting in 6.4 GB/s bandwidth shared by up to four processors. The Server I/O Hub (SIOH) provides connectivity between I/O bridge components to the node. The SPS has six identical ports supporting 3.2GB/s in each direction. The E8870 chipset supports both the IA32 and the IA64 processor family.

The actual configuration of the Dell PowerEdge 3250 server contains just two 1.3GHz Itanium-II (Madison) processors. Each processor has 16KB L1 Data and 16KB L1 Instruction cache, a 256KB 8 way set-associative L2 cache, and a 3MB 24-way set-associative L3 cache. The node contains a single SNC and a single SIOH with no SPS switches.

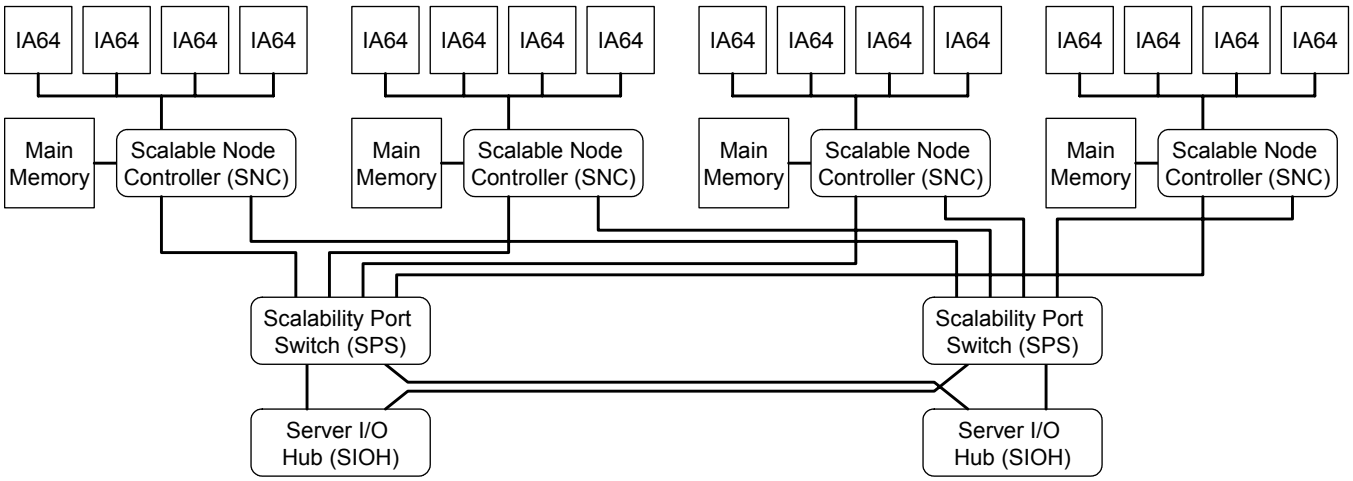

Figure 2. The configuration of a possible 16-way Itanium-II node using the Intel E8870 chipset.

3. LOW-LEVEL PERFORMANCE CHARACTERISTICS

Two micro-benchmarks were used to measure the achievable latency to memory and bandwidth.

1) Memory Latency - the latency to memory was measured using Memtime – a micro-benchmark in which one word per cache-line size (128 bytes) is accessed across the memory space in succession. By altering the size of the data memory available in Memtime, the latency to the different levels of the memory hierarchy can be measured.

2) Memory Bandwidth – the achievable bandwidth to memory was measured using Cachebench [7]. This micro-benchmark measures the bandwidth to the different levels in the memory hierarchy. It can be used to measures several memory characteristics: read only, write only, read/modify/write (an increment), memcpy, and memset.

Results from both micro-benchmarks are included below.

3.1 Memory Latency

The measured latency to all the available memory modules in both processing nodes is shown in Figure 3. Only the latency to the main memory is shown as the latency to the different cache levels is not dependent on the memory modules but is a function of the cache configuration and processor clock speed. Note that the Xeon node uses the left hand Y-axis and the Itanium-II node uses the right hand axis.

It can be seen that the measured latency on the Xeon node varies dependent upon the memory module. The slowest at 165 cycles is actually that which is rated the fastest (the Infineon CL2.0-2-2/TSOP). The Nanya CL2.5-3-3 is rated one of the two slowest but has a slightly lower latency at 160 cycles than the remaining four modules.

On the Itanium-II node the picture is a little different. Both the Infineon CL2.0-3-3/TSOP and the Micron have a latency of 261 cycles in comparison to the other four. The differences in the latency of the memory modules are not seen in the observed latency to memory – the increased complexity of the Itanium-II memory system effectively hides these differences. This complexity can also be noted that the actual number of cycles to memory on the Itanium is much higher than that on the Xeon node. However, in typical use on the Itanium-II, the compiler does a good job at hiding the effective latency – i.e. the latency that is actually experienced within a particular application.

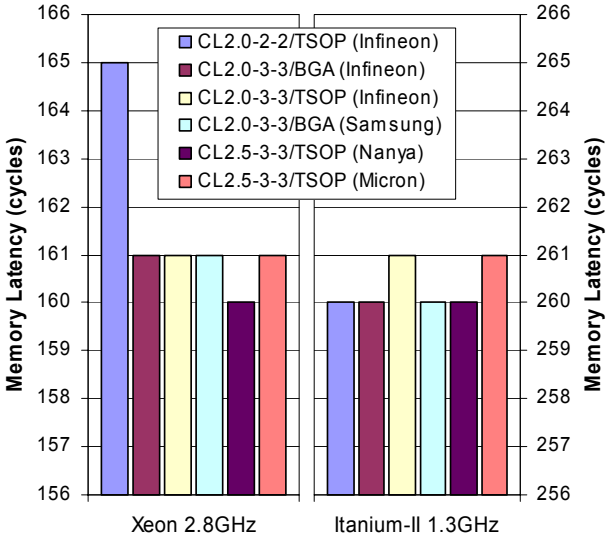

Figure 3. Measured latency to main memory on both processing nodes.

3.2 Memory Bandwidth

Cachebench measures several memory characteristics: read only, write only, read/modify/write (an increment), memcpy, and memset. A data array is used whose total memory footprint can be varied so as to expose the different levels in the memory hierarchy. A loop is used to provide a stride one access into the data array and thus produces a stream type access pattern. Additionally, the read only, write only and read/modify/write measurements can be made through a tuned option in which the

loops are manually unrolled by a factor of 8. This may or may not improve reported bandwidths as the effectiveness of the unrolling is heavily dependent on the compiler.

The main measurement loop in Cachebench can be considered to be:

```
for  (index = 0; index < limit; index++) {
    <Memory Operation>
}
```

where the individual memory operations are listed in Table 3 below.

Table 3. Measured Memory Operations

Memory Operation	Code
Read only	sum += x[index];
Write only	x[index] = wval;
Read/Modify/Write	x[index]++;
Memcpy	memcpy(x,y,bytes);
Memset	memset(x,0xf0,bytes);

An example of the read bandwidth as measured from Cachebench is shown in Figure 4 for the Infineon CL2.0-2-2 memory module. The bandwidth is shown for both processing nodes in two configurations. The first is when only a single processor is performing the memory read operation (the second processor being idle), and the second when both processors are performing memory reads. There is clearly a decrease in measured bandwidth per processor when both processors are performing memory operations due to contention on the front-side-bus. This can be significant and can reduce overall node performance. The measured 1-processor bandwidth is 1.75GB/s and the 2-processor bandwidth is 1.27GB/s (per processor) on the Xeon node. On the Itanium-II node, the measured 1-processor achieves 5.27GB/s and 2-processors achieve 2.47GB/s (per processor).

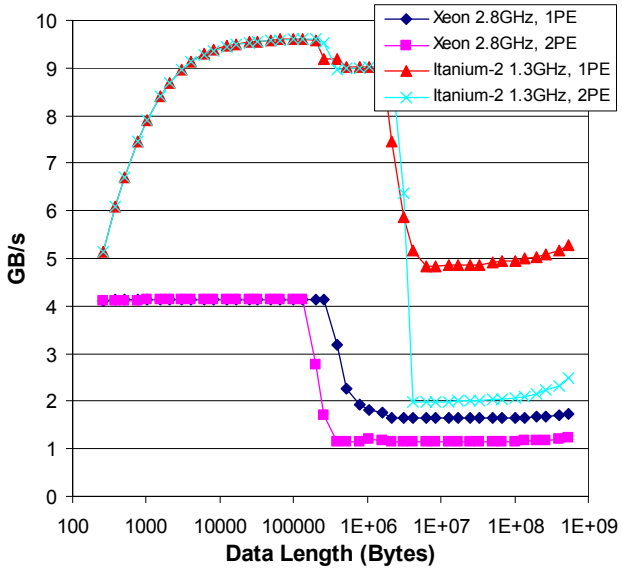

Figure 4. Measured read bandwidth on both processing nodes using the Infineon CL2.0-2-2 modules.

Table 4. Measured bandwidths (in GB/s) on the Infineon CL2.0-2-2 memory module.

	Xeon 2.8GHz		Itanium-II 1.3GHz	
	1PE	2PE	1PE	2PE
Read	1.663	1.148	5.460	3.329
ReadTuned	0.746	0.519	2.396	1.207
Write	1.462	1.135	4.701	2.246
WriteTuned	2.086	1.306	1.750	0.592
Rmw	0.737	0.566	1.184	0.707
RmwTuned	1.446	0.973	0.677	0.442
Memset	0.650	0.510	3.164	1.878
Memcpy	1.098	0.680	1.865	0.960

The caches sizes in the two processing nodes can also be seen in Figure 4. On Xeon Node, the L2 cache is 512KB, and on the Itanium-II the L2 Cache is 256KB, and L3 is 3MB. The increased bandwidth on the Itanium-II node is due to an increased width to memory.

The achievable bandwidth measured by Cachebench for all the operations listed in Table 3 as well as the tuned variants is included in Table 4. The bandwidth is listed for both processing nodes when using either 1 or 2 processors for the Infineon CL2.0-2-2 memory module. The bandwidths are reported in GB/s per processor (PE). The results are reported for the average bandwidths over the range of 8MB-200MB data array size which is considerably above the cache size.

Bandwidth measurements were also made for all the memory modules in both processing nodes. The differences between the memory modules is better seen by considering the relative performance for each module using the Infineon CL2.0-2-2 bandwidth as listed in Table 4 as a baseline. The relative performance of the other memory modules is best seen graphically in Figure 5 for the Xeon node and Figure 6 for the Itanium-II node.

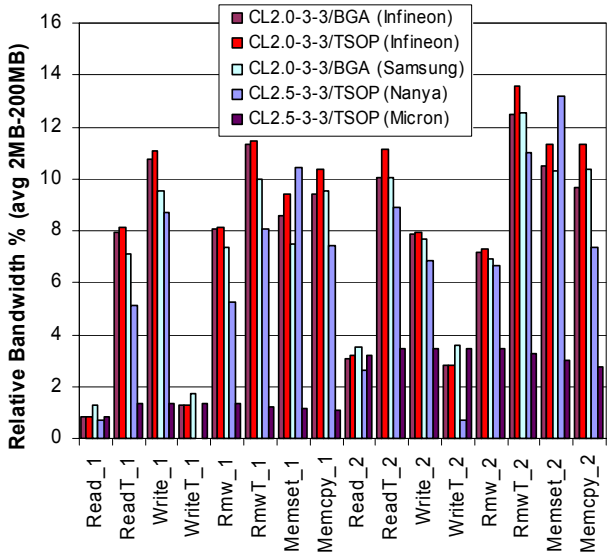

Figure 5. Relative read bandwidth performance of memory modules 2 to 6 in comparison to module 1 on the Xeon node.

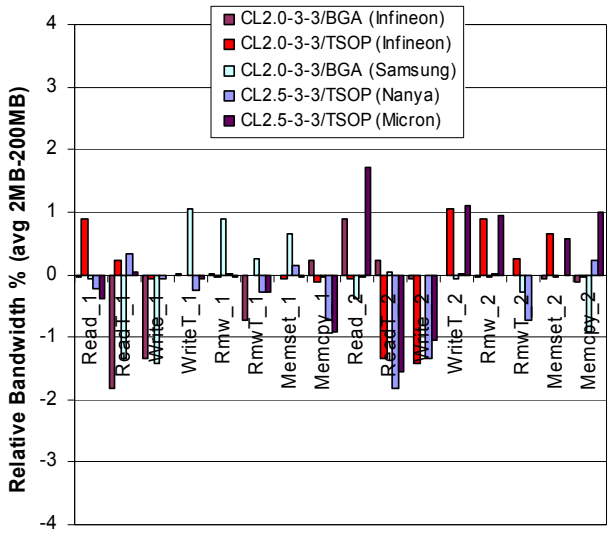

Figure 6. Relative read performance of memory modules 2 to 6 in comparison to module 1on the Itanium-II node.

It can be seen that there is a difference in the relative performance between the memory modules on the two processing nodes. On the Xeon node, the relative performance difference is between 1% and 14% across all the cases measured. In all cases the relative bandwidth of the modules is positive indicating a higher performance than the Infineon CL2.0-2-2 module. However, the picture is different on the Itanium-II node. Firstly, the relative performance ranges from -1.8% up to +1.7%. There is no clear benefit from using a particular memory module over another in this processing node. The impact of the differences in the performance of the memory sub-system on the Xeon node will impact the performance of applications. The extent of the impact cannot be determined from Figure 5 (or Figure 3) as the memory access pattern of an application will not match that measured in either of the latency or the bandwidth benchmarks. In Section 4 below, the performance of two applications are analyzed.

4. APPLICATION PERFORMANCE

Two applications are used to analyze the performance of all the memory modules on the two processing nodes. These are: Sweep3D and SAGE. Both of these applications are representative of significant elements of the ASC workload which are executed on many of the largest supercomputers including all ASC machines.

4.1 Sweep3D

Sweep3D is a time-independent, Cartesian-grid, single-group, "discrete ordinates" deterministic particle transport code. Estimates indicate that deterministic particle transport accounts for 50-80% of the execution time of many realistic simulations on current ASC systems. The basis for neutron transport simulation is the time-independent, multigroup, inhomogeneous Boltzmann transport equation.

The performance given by Sweep3D is representative of larger ASC applications but the processing is solved on a reduced number of unknowns. Sweep3D uses a 3-dimensional spatial grid which is partitioned in 2-dimensions for high performance, parallel processing. A sub-grid on a single processor comprises I x J x K cells. The number of cells per processor in both I and J are varied in this testing (from 5 to 15) with the number of cells in the K dimension fixed at 400. This results in the number of cells per processor varying from 5x5x400 = 10,000 cells up to 14x14x400 = 78,400 cells. The processing typically scans this 3-D sub-grid volume in three nested loops for the K, J and I (innermost) dimensions respectively. The main data arrays are directly indexed in this loop ordering.

The performance was measured on all memory modules for both a single processor execution and a dual-processor execution (as a parallel MPI job communicating via shared memory). Note that each measurement was taken from a runtime of several minutes to reduce measurement noise. Figure 7 shows the performance of Sweep3D on both processing nodes for the 1-processor and 2-processor executions using the Infineon CL2.0-2-2 memory modules. The scaling of the problem sub-grid per processor is on the X-axis, and the grind time is on the Y-axis. The grind time here is the time taken to process a single cell – the overall run-time of Sweep3D is the grind time multiplied by both the sub-grid volume and the number of iterations required for convergence in the calculation. A lower grind time indicates a higher performance. It can be seen that a higher performance is achieved on the Itanium-II node than on the Xeon node.

The relative performance on Sweep3D of the other 5 memory modules is shown in Figure 8 for both processing nodes. The relative performance represents an average over the sub-grid problem sizes considered in Figure 7. On the Xeon node, there is a difference in performance across memory modules. The highest performance is achieved on the Infineon CL2.0-3-3/TSOP module achieving 6.77% better performance on 2-processors than the Infineon CL2.0-2-2 module. The relative performance differences on the Itanium-II node are marginal. The Nanya CL2.5-3-3 is 1% slower on 2-processors than the Infineon CL2.0-2-2 module.

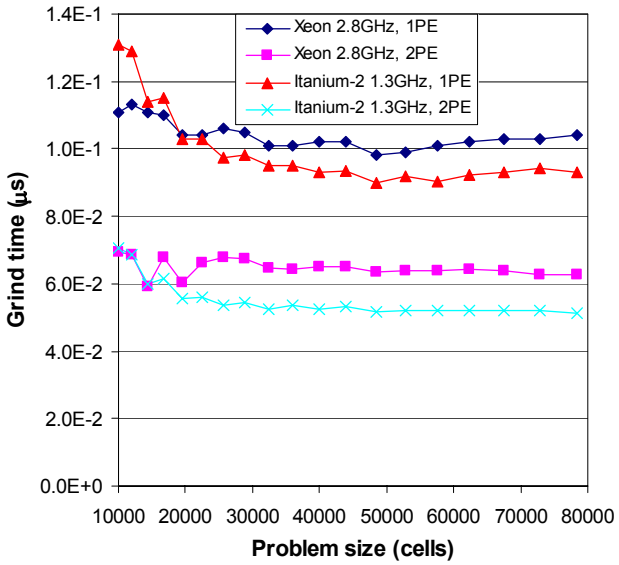

Figure 7. Performance of Sweep3D on both processing nodes using the Infineon CL2.0-2-2 modules.

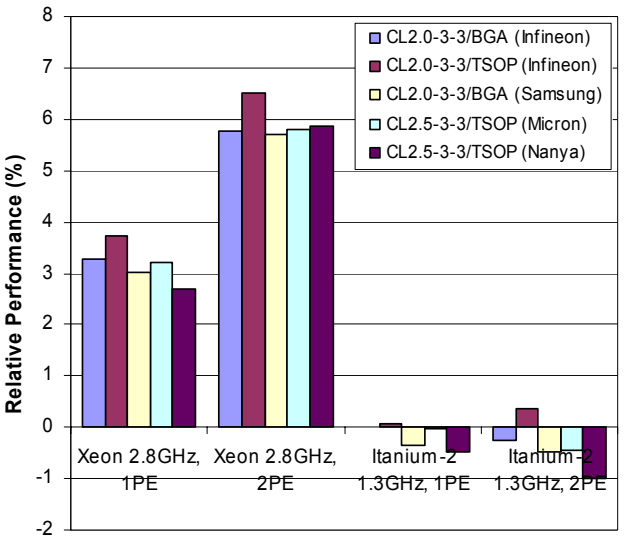

Figure 8. Relative Sweep3D performance on memory modules 2 to 6 in comparison to module 1.

4.2 SAGE

SAGE (SAIC's Adaptive Grid Eulerian hydrocode) is a multidimensional (1D, 2D, and 3D), multi-material, Eulerian hydrodynamics code with adaptive mesh refinement (AMR). SAGE performs hydro-dynamic and heat radiation operations on a structured spatial mesh that can be adaptively refined on a cell by cell basis as necessary at the end of each processing cycle. Each cell at the topmost level (level 0) can be considered as the root node of an oct-tree of cells in lower levels.

Three input decks are used for SAGE to examine its performance under different conditions. These are:

1) timing_a – hydro (no heat) using an AMR grid
2) timing_b – hydro and heat using an AMR grid
3) timing_h – hydro and heat on a fixed grid (no AMR)

Both timing_a and timing_b performs AMR operations at the end of each processing cycle and thus the number of cells per processor varies as the calculation progresses. The use of both hydro and heat is more representative of the actual use of SAGE. The timing_h input deck does not perform any AMR operation and thus the number of cells per processor is fixed throughout a particular execution. This enables an accurate study of varying the number of cells per processor without the added complication of the AMR process.

The computation within SAGE requires a lot of gather-scatter type memory operations. That is, one or more variable is typically gathered via an indirection array into a contiguous data array, processed, and then the results scattered to their destination location via the same indirection array. This type of operation is necessary to support the AMR operation.

For each of the three input decks, the performance was measured for both 1- and 2-processor executions on both processing nodes using all the available memory modules. Note again that each measurement was taken from a runtime of several minutes to reduce measurement noise. The measured performance using the Infineon CL2.0-2-2 memory modules is shown in Figures 9, 10, and 11 for the three input decks. In Figures 9 and 10, the performance per iteration cycle was recorded since both the timing_a and timing_b input decks perform adaptation, and thus alter the problem size per processor across iterations. This can vary the achieved performance on an iteration basis. In Figure 11, the performance per problem size is shown when using the timing_h input deck. This was obtained for a 10 iteration processing run with the problem size constant across iterations. The performance metric that is used is the number of cell-updates-per-second-per-processor (cc/s/pe). This is a rate based metric, and thus a higher value represents a higher performance.

The relative performance on SAGE on the remaining 5 memory modules is shown in Figures 12, 13, and 14 for the three input decks. In each case the performance is shown for 1- and 2-processor executions on both processing nodes.

The results are very consistent. In the case of the Itanium-II node, there is no significant difference across all memory modules and tests performed. In the case of the Xeon node, the highest performance is achieved on the Infineon CL2.0-3-3/TSOP (module 3) in all cases. The performance of this is at most 18.9% higher than the Infineon CL2.0-2-2 module.

5. ANALYSIS OF RESULTS

A summary of the relative performance of memory modules 2-6 in comparison to memory module 1 is listed in Table 5 for Sweep3D and SAGE as well as the observed memory read and memory write bandwidths. This is shown for the two processing nodes when using 1-processor or 2-processors in the node. The Sweep3D and SAGE performances shown in Table 5 are an average over all the testing performed for each of the applications.

The results are quite consistent across the testing. The primary factor that should be noted is that the memory behaved differently in the two processing nodes. Very little difference has been noted on the Itanium-II node across all the testing. Whereas, almost a 20% difference was noted on the Xeon node for some of the SAGE testing. This observation in itself was not expected and is attributed to the very different memory systems between the two processors. The Itanium-II processors have a higher latency to main memory but rely heavily on compiler technology to hide this latency. The effectiveness of this latency hiding needs to be looked at further. An interesting memory model [6] which uses the processor performance counters to quantify the causes of memory stalls within the processor is currently being applied to explore these issues.

There are quite large differences in the observed performances on the Xeon node. From the application perspective the difference in performance ranges from 2.68% to 6.51% across all the test cases on Sweep3D, and from 10.4% to 18.9% across all the test cases on SAGE. The exact performance differences are dependant on the application being executed, and also on the exact problem characteristics being processed.

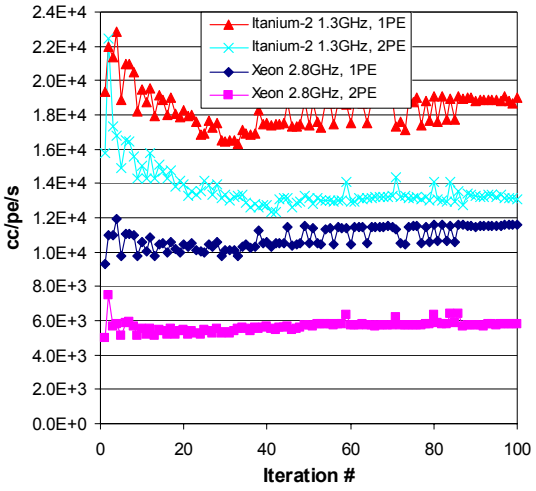

Figure 9. Performance of SAGE (timing_a) on both processing nodes using the Infineon CL2.0-2-2 modules.

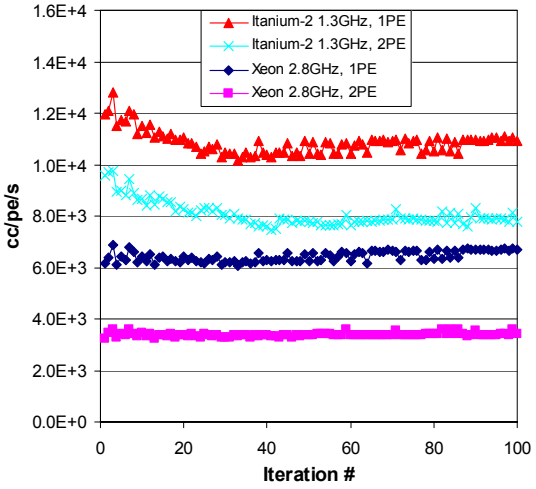

Figure 10. Performance of SAGE (timing_b) on both processing nodes using the Infineon CL2.0-2-2 modules.

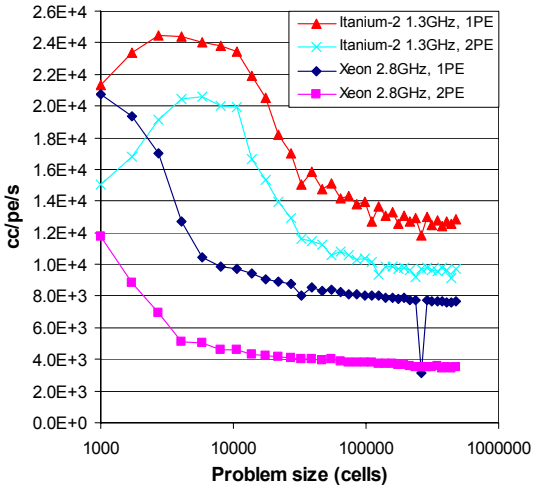

Figure 11. Performance of SAGE (timing_h) on both processing nodes using the Infineon CL2.0-2-2 modules.

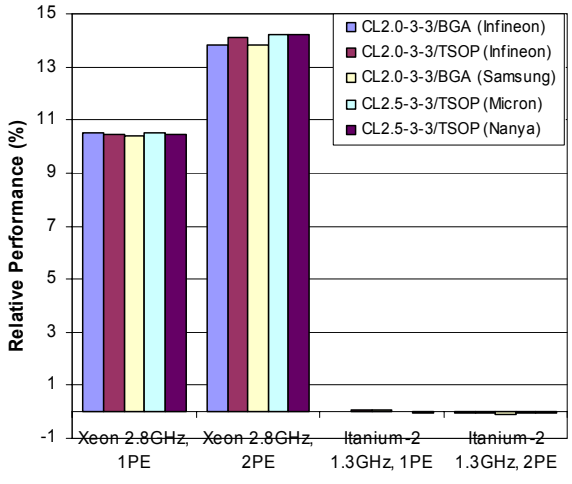

Figure 12. Relative SAGE (timing_a) performance on memory modules 2 to 6 in comparison to module 1.

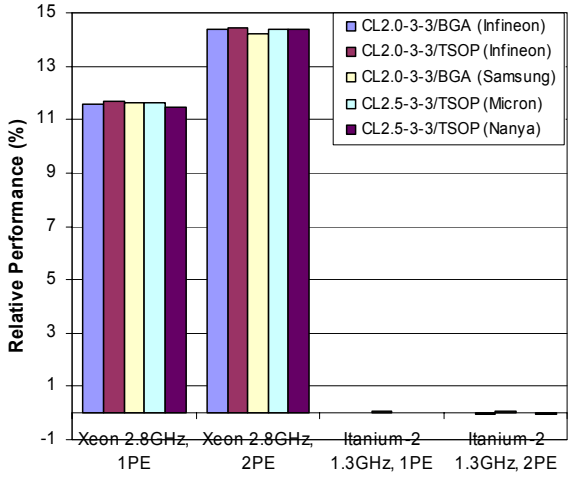

Figure 13. Relative SAGE (timing_b) performance on memory modules 2 to 6 in comparison to module 1.

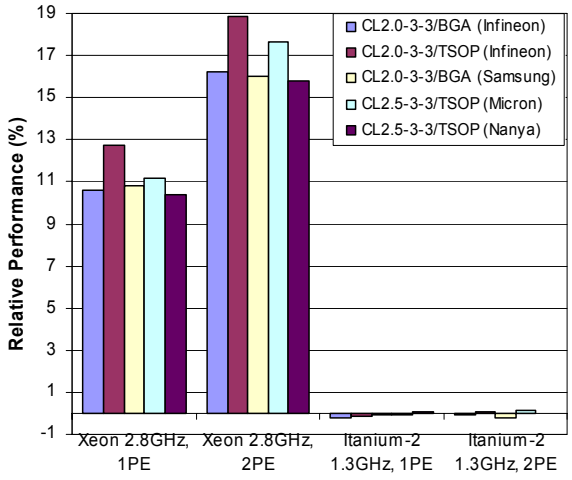

Figure 14. Relative SAGE (timing_h) performance on memory modules 2 to 6 in comparison to module 1.

Table 5. Summary of the relative performance of all memory modules.

		Xeon					Itanium- II				
	Module	2	3	4	5	6	2	3	4	5	6
1 PE	Read Bandwidth	0.80%	0.80%	1.29%	0.73%	0.84%	-0.04%	0.90%	-0.08%	-0.23%	-0.38%
	Write Bandwidth	10..8%	11.1%	9.52%	8.70%	1.31%	-1.35%	-0.06%	-1.42%	-1.35%	-0.01%
	Sweep3D	3.28%	3.73%	3.02%	2.68%	3.20%	0.01%	0.01%	-0.3%	-0.5%	-0.02%
	SAGE	10.9%	11.6%	11.0%	10.8%	11.1%	-0.05%	-0.04%	0.05%	0.02%	-0.02%
2 PEs	Read Bandwidth	3.05%	3.22%	3.55%	2.65%	3.21%	0.90%	-0.08%	-0.38%	-0.04%	0.90%
	Write Bandwidth	7.86%	7.93%	7.70%	6.88%	3.46%	-0.06%	-1.42%	-0.1%	-1.35%	-1.0%
	Sweep3D	5.78%	6.51%	5.70%	5.86%	5.79%	-0.25%	0.37%	-0.48%	-1.01%	-0.45%
	SAGE	14.8%	15.8%	14.7%	14.8%	15.4%	-0.04%	-0.03%	-0.08%	-0.03%	0.03%

6. DISCUSSION

The initial expectations for this work were that the micro-benchmarks and the applications would see a small percentage difference in performance based on the specifications of the memory modules used. The results however, were non-intuitive. Almost no difference was perceivable in the Itanium system and one memory module on the Xeon based system showed almost a 20% slowdown. Furthermore, the module with the worst performance was the one with the highest performance specifications.

To investigate these issues a closer look must be given to the way in which the chipset interfaces to the memory modules. Each memory module has a Serial Presence Detect (SPD) EEPROM. The SPD contains the timing information specific to the memory module. Unfortunately, this information is not always used by the motherboard chipset. The chipsets typically contain a Memory Hub Controller (MHC). The registers of the MHC are used to control the timing to the memory modules. To verify that the SPD information is being correctly used, the settings of the MHC registers need to be examined.

In the case of the Xeon's GC-LE chipset an Intel proprietary tool was used to verify, and to modify these timing settings. It was found that the memory timing control register (MTCR) of the MHC was not set to the correct SPD values. For memory module 1, the latencies were actually set to be 1 clock cycle longer for both tRCD and tRP within the MTCR. This indicates a chipset idiosyncrasy such that the latency values within the MTCR were lower than they should be for the slower modules and higher for the faster modules. After the values were manually corrected for memory 1, the performance was re-measured on Cachebench and the two application codes. The performance improvement observed is listed in Table 6. This effectively results in memory module 1 achieving the same performance as the other 5 modules.

Table 6. Performance gain of memory module 1 with corrected MTCR values.

	% Gain 1-PE	% Gain 2-PE
Read Bandwidth	1.3	4.3
Sweep3D	3.4	5.9
SAGE	11.5	13.3

In the Itanium's case it is assumed that the large latency of the E8870 masks any perceivable performance difference. This is due to the expandability of the chipset that can scale up to 16 processors. Though it is possible that the SPD information is being ignored and the different memory modules were all accessed using the same timing values. This has yet to be verified.

To generalize the results, the specifications of the memory modules themselves have less impact on achieved performance then the interface between the chipset and the modules. Also it should be noted that in the case of the Xeon node all the memory modules performed at the faster latency settings regardless of their individual specifications. Validating memory with chipsets is seen as a crucial step in configuration of compute nodes. Significant performance can be lost when the configuration of the chipset does not accurately match those of the memory modules. These setting are not always accessible to the end-user and so one must rely on the node supplier. The cost of including memory modules with higher performance specifications may not be justified as an increase in achievable application performance may not be realized.

7. SUMMARY

Memory is a significant part of the construction of commodity based high performance systems. However, the performance of the memory is often overlooked and rather just its capacity and rated bus speed is specified. Memory is very much a commodity in its own right and many alternatives are available for use in any given processing node. Memory modules vary in many characteristics including: latency, bandwidth, packaging, and manufacturer. In this work we have explored the performance of six different types of memory modules in two very different server nodes – a dual Xeon IA32 2.8GHz node, and a dual Itanium-II 1.3GHz node. The aim of this work was to see if there was an identifiable performance differences across these memory modules and processing nodes.

The testing performed included examining low-level performance characteristics – those of the latency to memory and the achievable bandwidth to memory from the processors within the node. In addition the performance of two applications that are representative of significant components of the ASC workload have been examined. Although a quantitative performance analysis of these two applications on the six memory modules has

been provided, a qualitative feel for the performance differences is also given. This results in the overall conclusion that the choice of the memory can affect performance significantly in some situations, but not in all. Applications have different memory characteristics and do not necessarily directly reflect just the memory latency and bandwidths.

In particular this analysis has shown that the choice of memory for the Itanium-II node is not a significant factor that will affect the overall performance of the processing node. The greatest observed performance difference was 1% - this was deemed insignificant when the multitude of tests were considered in combination. On the other hand, the performance difference observed on the Xeon processing node did vary by between 2.7% and 18.9% depending on the actual application / input being processed. After an in-depth investigation it was found that this difference was due to a miss-configuration within the chip-set rather than to the differences in the memory performance attributes.

This work is being extended to include the examination of a dual Opteron node, a dual Intel x86-64 Nocona node, and a dual PPC-G5 node. In addition the performance differences across the memory modules will be further explored by examination of the memory stall times and the effective memory latency that is being achieved in the applications [6]. Scaling issues are also a concern on large-scale supercomputers. These will be examined by applying accurate performance models of the applications developed at Los Alamos. The effects of memory performance differences will be investigated on applications scaling to thousands of processors.

Memory is a commodity and its performance examined in order to optimize the performance of a processing node.

"The difference between false memories and true ones is the same as for jewels: it is always the false ones that look the most real, the most brilliant."

– Salvador Dali

ACKNOWLEDGEMENTS

We would like to thank Hossein Amidi and Sat Kolli of Smart Modular Technologies for discussions on this work and for supplying diagnostic tools. Los Alamos National Laboratory is operated by the University of California for the U.S. Department of Energy under contract W-7405-ENG-36.

REFERENCES

[1] Hoisie A, Lubeck O, Wasserman H.J. Performance and Scalability Analysis of Teraflop-Scale Parallel Architectures using Multidimensional Wavefront Applications. *Int. J. of High Performance Computing Applications*, **14**(4): 330-346.

[2] Intel E8870 Scalable Node Controller (SNC) Datasheet, Document Number: 251112-001, August 2002. Availble from: ftp://download.intel.com/design/chipsets/datashts/25111203.pdf

[3] Intel PC SDRAM Serial Presence Detect (SPD) Specification, Revision 1.2A, dec, 1997, available from: http://www.intel.com/design/chipsets/memory/spdsd12a.pdf

[4] Kerbyson D J, Alme H J, Hoisie A, Petrini F, Wasserman H J, Gittings M. Predictive Performance and Scalability Modeling of a Large-Scale Application. *Proceedings SC'01*, Denver, November 10-16, 2001.

[5] Kerbyson D J, Hoisie A, Wasserman H J. A Comparison Between the Earth Simulator and AlphaServer Systems using Predictive Application Performance Models. *Proceedings Int. Parallel and Distributed Processing Symposium (IPDPS),* Nice, France, April 21-25, 2003.

[6] Lubeck O., Luo Y., Wasserman H.J., Bassetti F. An empirical hierarchical memory model based on hardware performance counters. *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications*, July 1998.

[7] Mucci, P, Cachebench software, available from: http://icl.cs.utk.edu/projects/llcbench/cachebench.html